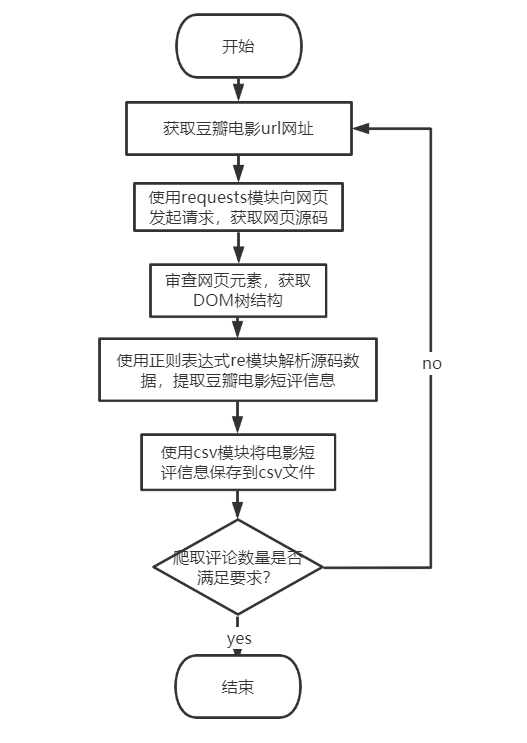
详细设计

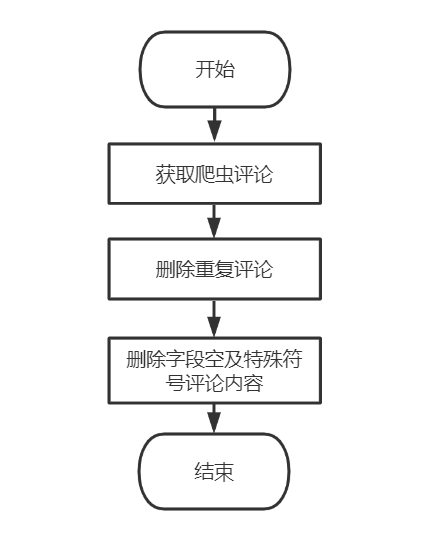
#### 1.数据获取模块

首先获取豆瓣网站电影相关的url地址，使用requests模块向网页发起请求，获取网页源码。审查网页元素，获取dom树结构，再利用正则表达式匹配我们想要的comment域，提取豆瓣电影的短评信息，由于豆瓣每页的评论数是20，所以我们每部电影爬取100页，根据url地址的limit后缀让程序自动换页并将这些评论存入csv文件。当评论总数达到要求时，将他们改成txt格式。



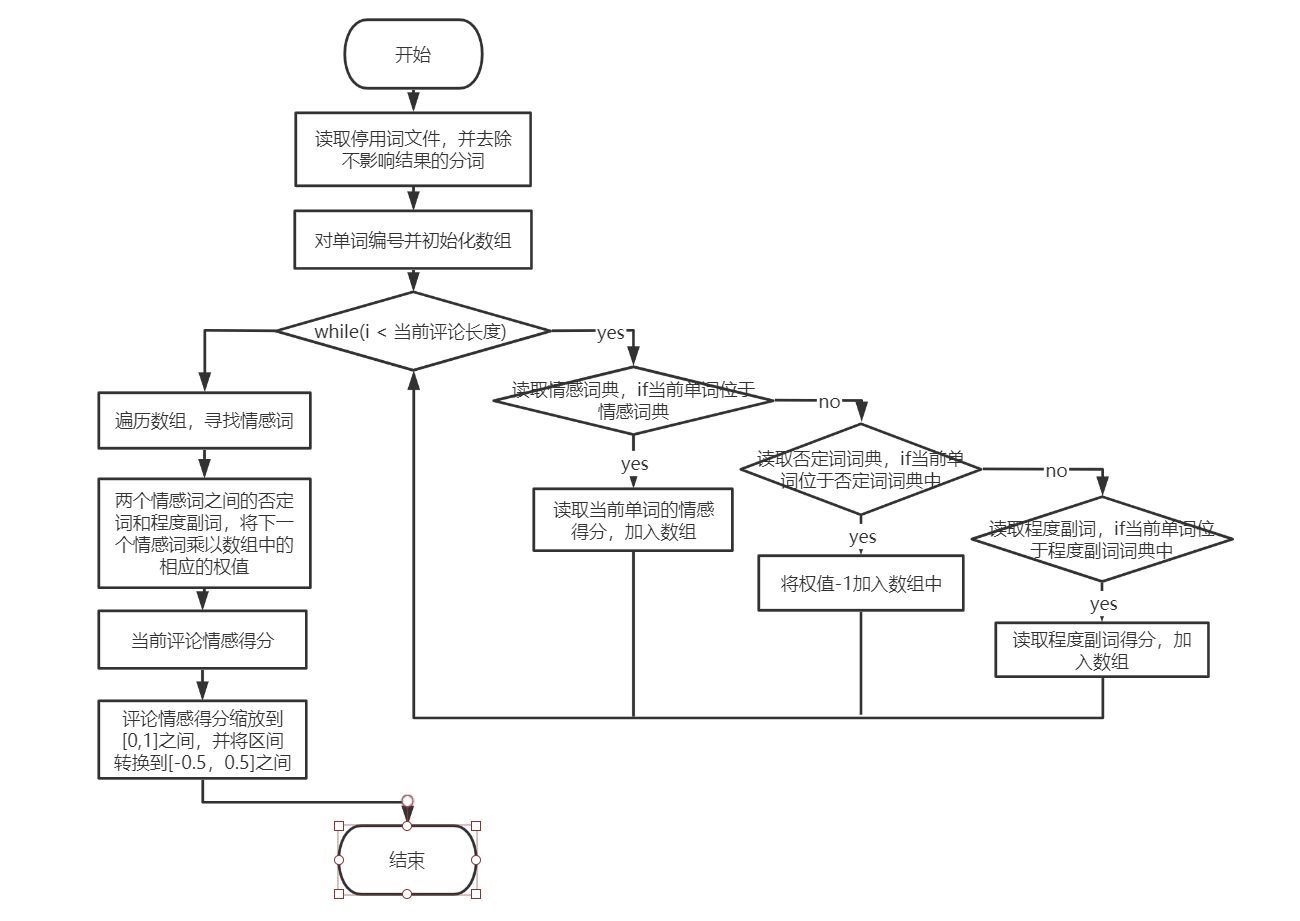
#### 2.数据预处理模块

将txt格式的数据审查，检查到哪一行和上面一行一致或者字段为空，就将该数据视为含缺失值，并将其删除。



#### 3.情感分析模块

首先将每条评论进行jieba分词，将每段评论分成单独的词，然后调用我们手动添加的停用词词典，将一些停用词删除，删除之后根据bosonNLP词典定位情感词出现的位置和它的分数，对于每个情感词，其前置的否定词和程度副词可以影响它的得分。查询我们手动添加的否定词和程度副词，否定词出现，就将它乘以-1，程度副词出现，就将它乘以相应的权值，最后汇总情感词的得分，即得出单条评论的得分，区间为0~1。对每部电影每条评论进行打分，该阶段完成。



#### 4.可视化模块

我们将每部电影的评分进行可视化，以影评情感得分为x轴，影评数量为y轴。生成相应的柱状图。

还可以对每部电影根据分词和去停用词之后生成相应的词云图。